On the Global Optimality of Model-Agnostic Meta-Learning: Reinforcement Learning and Supervised Learning

Lingxiao Wang [1] **Qi Cai** [1] **Zhuoyan Yang** [2] **Zhaoran Wang** [1]

Abstract

Model-agnostic meta-learning (MAML) formulates meta-learning as a bilevel optimization problem, where the inner level solves each subtask based on a shared prior, while the outer level searches for the optimal shared prior by optimizing its aggregated performance over all the subtasks. Despite its empirical success, MAML remains less understood in theory, especially in terms of its global optimality, due to the nonconvexity of the meta-objective (the outer-level objective). To bridge such a gap between theory and practice, we characterize the optimality gap of the stationary points attained by MAML for both reinforcement learning and supervised learning, where the inner-level and outer-level problems are solved via first-order optimization methods. In particular, our characterization connects the optimality gap of such stationary points with (i) the functional geometry of inner-level objectives and (ii) the representation power of function approximators, including linear models and neural networks. To the best of our knowledge, our analysis establishes the global optimality of MAML with nonconvex meta-objectives for the first time. [1]

1. Introduction

Meta-learning aims to find a prior that efficiently adapts to a new subtask based on past subtasks. One of the most popular meta-learning methods, namely model-agnostic meta-learning (MAML) (Finn et al., 2017a), is based on bilevel optimization, where the inner level solves each subtask based on a shared prior, while the outer level optimizes the aggregated performance of the shared prior over all the subtasks. In particular, MAML associates the solution to each subtask with the shared prior through one step of gradient descent based on the subtask data. Due to its model-agnostic property, MAML is widely adopted in reinforcement learning (Finn et al., 2017a;b; Xu et al., 2018; Nagabandi et al., 2018; Gupta et al., 2018; Yu et al., 2018; Mendonca et al., 2019) and supervised learning (Finn et al., 2017a; Li et al., 2017; Finn et al., 2018; Rakelly et al., 2018; Yoon et al., 2018).

Despite its popularity in empirical studies, MAML is scarcely explored theoretically. In terms of the global optimality of MAML, (Finn et al., 2019) show that the meta-objective is strongly convex assuming that the inner-level objective is strongly convex (in its finite-dimensional parameter). However, such an assumption fails to hold for neural function approximators, which leads to a gap between theory and practice. For nonconvex meta-objectives, (Fallah et al., 2019) characterize the convergence of MAML to a stationary point under certain regularity conditions. Meanwhile, (Rajeswaran et al., 2019) propose a variant of MAML that utilizes implicit gradients, which is also guaranteed to converge to a stationary point. However, the global optimality of such stationary points remains unclear. On the other hand, (Pentina & Lampert, 2014; Amit & Meir, 2017) establish PAC-Bayes bounds for the generalization error of two variants of MAML. However, such generalization guarantees only apply to the global optima of the two meta-objectives rather than their stationary points.

In this work, we characterize the global optimality of the ϵ-stationary points attained by MAML for both reinforcement learning (RL) and supervised learning (SL). For meta-RL, we study a variant of MAML, which associates the solution to each subtask with the shared prior, namely π_θ, through one step of proximal policy optimization (PPO) (Schulman et al., 2015; 2017) in the inner level of optimization. In the outer level of optimization, we maximize the expected total reward associated with the shared prior aggregated over all the subtasks. We prove that the ϵ-stationary point attained by such an algorithm is (approximately) globally optimal given that the function approximator has sufficient

[1]Department of Industrial Engineering and Management Sciences, Northwestern University, USA [2]Department of Operations Research and Financial Engineering, Princeton University, USA. Correspondence to: Lingxiao Wang <lingxiaowang2022@u.northwestern.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

[1]See https://arxiv.org/abs/2006.13182 for the full version.

representation power. For example, for the linear function approximator $\pi_\theta(s,a) \propto \exp(\phi(s,a)^\top\theta)$, the optimality gap of the ϵ-stationary point is characterized by the representation power of the linear class $\{\phi(\cdot,\cdot)^\top v : v \in \mathcal{B}\}$, where $\mathcal{B}$ is the parameter space (which is specified later). The core of our analysis is the functional one-point monotonicity (Facchinei & Pang, 2007) of the expected total reward $J(\pi)$ with respect to the policy π (Liu et al., 2019) for each subtask. Based on a similar notion of functional geometry in the inner level of optimization, we establish similar results on the optimality gap of meta-SL. Moreover, our analysis of both meta-RL and meta-SL allows for neural function approximators. More specifically, we prove that the optimality gap of the attained ϵ-stationary points is characterized by the representation power of the corresponding classes of overparameterized two-layer neural networks.

Challenge. We highlight that the bilevel structure of MAML makes it challenging for the analysis of its global optimality. In the simple case where the inner-level objective is strongly convex and smooth, (Finn et al., 2019) show that the meta-objective is also strongly convex assuming that the stepsize of inner-level optimization is sufficiently small.

- In practice, however, both the inner-level objective and the meta-objective can be nonconvex, which leads to a gap between theory and practice. For example, the inner-level objective of meta-RL is nonconvex even in the (infinite-dimensional) functional space of policies.
- Even assuming that the inner-level objective is convex in the (infinite-dimensional) functional space, nonlinear function approximators, such as neural networks, can make the inner-level objective nonconvex in the finite-dimensional space of parameters.
- Furthermore, even for linear function approximators, the bilevel structure of MAML can make the meta-objective nonconvex in the finite-dimensional space of parameters, especially when the stepsize of inner-level optimization is large.

In this work, we tackle all these challenges by analyzing the global optimality of both meta-RL and meta-SL for both linear and neural function approximators.

Contribution. Our contribution is three-fold. First, we propose a meta-RL algorithm and characterize the optimality gap of the ϵ-stationary point attained by such an algorithm for linear function approximators. Second, under an assumption on the functional convexity of the inner-level objective, we characterize the optimality gap of the ϵ-stationary point attained by meta-SL. Finally, we extend our optimality analysis for linear function approximators to handle overparameterized two-layer neural networks. To the best of our knowledge, our analysis establishes the global optimality of MAML with nonconvex meta-objectives for the first time.

Related Work. Meta-learning is studied by various communities (Evgeniou & Pontil, 2004; Thrun & Pratt, 2012; Pentina & Lampert, 2014; Amit & Meir, 2017; Nichol et al., 2018; Nichol & Schulman, 2018; Khodak et al., 2019). See (Pan & Yang, 2009; Weiss et al., 2016) for the surveys of meta-learning and (Taylor & Stone, 2009) for a survey of meta-RL. Our work focuses on the model-agnostic formulation of meta-learning (MAML) proposed by (Finn et al., 2017a). In contrast to existing empirical studies, the theoretical analysis of MAML is relatively scarce. (Fallah et al., 2019) establish the convergence of three variants of MAML for nonconvex meta-objectives. (Rajeswaran et al., 2019) propose a variant of MAML that utilizes implicit gradients of the inner level of optimization and establish the convergence of such an algorithm. This line of work characterizes the convergence of MAML to the stationary points of the corresponding meta-objectives. Our work is complementary to this line of work in the sense that we characterize the global optimality of the stationary points attained by MAML. Meanwhile, (Finn et al., 2019) propose an online algorithm for MAML with regret guarantees, which rely on the strong convexity of the meta-objectives. In contrast, our work tackles nonconvex meta-objectives, which allows for neural function approximators, and characterizes the global optimality of MAML. (Mendonca et al., 2019) propose a meta-policy search method and characterize the global optimality for solving the subtasks under the assumption that the meta-objective is (approximately) globally optimal. Our work is complementary to their work in the sense that we characterize the global optimality of MAML in terms of optimizing the meta-objective. See also the concurrent work (Wang et al., 2020).

There is a large body of literature that studies the training and generalization of overparameterized neural networks for SL (Daniely, 2017; Jacot et al., 2018; Wu et al., 2018; Allen-Zhu et al., 2018a;b; Du et al., 2018a;b; Zou et al., 2018; Chizat & Bach, 2018; Li & Liang, 2018; Cao & Gu, 2019a;b; Arora et al., 2019; Lee et al., 2019; Bai & Lee, 2019). See (Fan et al., 2019) for a survey. In comparison, we study MAML with overparameterized neural networks for both RL and SL. The bilevel structure of MAML makes our analysis significantly more challenging than that of RL and SL.

Notation. We denote by $[n] = \{1, 2, ..., n\}$ the index set. Also, we denote by $x = ([x]_1^\top, \ldots, [x]_m^\top)^\top \in \mathbb{R}^{md}$ a vector in $\mathbb{R}^{md}$, where $[x]_k \in \mathbb{R}^d$ is the k-th block of x for $k \in [m]$. For a real-valued function f defined on $\mathcal{X}$, we denote by $\|f(\cdot)\|_{p,\nu} = \{\int_{\mathcal{X}} f^p(x)\,\mathrm{d}\nu(x)\}^{1/p}$ the $L_p(\nu)$-norm of f, where ν is a measure on $\mathcal{X}$. We write $\|f(\cdot)\|_{2,\nu} = \|f(\cdot)\|_\nu$ for notational simplicity and $\|f\|_{p,\nu} = \|f(\cdot)\|_{p,\nu}$ when the

variable is clear from the context. For a vector $\phi \in \mathbb{R}^n$, we denote by $\|\phi\|_2$ the ℓ_2-norm of ϕ.

2. Background

In this section, we briefly introduce reinforcement learning and meta-learning.

2.1. Reinforcement Learning

We define a Markov decision process (MDP) by a tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma, \zeta)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, respectively, P is the Markov kernel, r is the reward function, which is possibly stochastic, $\gamma \in (0,1)$ is the discount factor, and ζ is the initial state distribution over $\mathcal{S}$. In the sequel, we assume that $\mathcal{A}$ is finite. An agent interacts with the environment as follows. At each step t, the agent observes the state s_t of the environment, takes the action a_t, and receives the reward $r(s_t, a_t)$. The environment then transits into the next state according to the distribution $P(\cdot \,|\, s_t, a_t)$ over $\mathcal{S}$. We define a policy π as a mapping from $\mathcal{S}$ to distributions over $\mathcal{A}$. Specifically, $\pi(a \,|\, s)$ gives the probability of taking the action a at the state s. Given a policy π, we define for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ the corresponding state- and action-value functions V^π and Q^π as follows,

$$V^\pi(s) = (1-\gamma) \cdot \mathbb{E}\bigg[\sum_{t=0}^{\infty} \gamma^t \cdot r(s_t, a_t) \,\bigg|\, s_0 = s\bigg], \tag{2.1}$$

$$Q^\pi(s,a) = (1-\gamma) \cdot \mathbb{E}\bigg[\sum_{t=0}^{\infty} \gamma^t \cdot r(s_t, a_t) \,\bigg|\, s_0 = s, a_0 = a\bigg], \tag{2.2}$$

where $s_{t+1} \sim P(\cdot \,|\, s_t, a_t)$ and $a_t \sim \pi(\cdot \,|\, s_t)$ for all $t \geq 0$. Correspondingly, the advantage function A^π is defined as follows,

$$A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s), \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}. \tag{2.3}$$

A policy π induces a state visitation measure ν_π on $\mathcal{S}$, which takes the form of

$$\nu_\pi(s) = (1-\gamma) \cdot \sum_{t=0}^{\infty} \gamma^t \cdot \mathbb{P}(s_t = s), \tag{2.4}$$

where $s_0 \sim \zeta$, $s_{t+1} \sim P(\cdot \,|\, s_t, a_t)$, and $a_t \sim \pi(\cdot \,|\, s_t)$ for all $t \geq 0$. Correspondingly, we define the state-action visitation measure by $\sigma_\pi(s,a) = \pi(a \,|\, s) \cdot \nu_\pi(s)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, which is a probability distribution over $\mathcal{S} \times \mathcal{A}$. The goal of reinforcement learning is to find the optimal policy π^* that maximizes the expected total reward $J(\pi)$, which is defined as

$$J(\pi) = \mathbb{E}_{s\sim\zeta}\big[V^\pi(s)\big] = \mathbb{E}_{(s,a)\sim\sigma_\pi}\big[r(s,a)\big]. \tag{2.5}$$

When $\mathcal{S}$ is continuous, maximizing $J(\pi)$ over all possible π is computationally intractable. A common alternative is to parameterize the policy by π_θ with the parameter $\theta \in \Theta$, where Θ is the parameter space, and maximize $J(\pi_\theta)$ over $\theta \in \Theta$.

2.2. Meta-Learning

In meta-learning, the meta-learner is given a sample of learning subtasks $\{\mathcal{T}_i\}_{i\in[n]}$ drawn independently from the task distribution ι and a set of parameterized algorithms $\mathcal{A} = \{\mathscr{A}_\theta : \theta \in \Theta\}$, where Θ is the parameter space. Specifically, given θ, the algorithm $\mathscr{A}_\theta \in \mathcal{A}$ maps from a learning subtask $\mathcal{T}$ to its desired outcome. For example, an algorithm that solves reinforcement learning subtasks maps from an MDP $\mathcal{T} = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \zeta)$ to a policy π, aiming at maximizing the expected total reward $J(\pi)$ defined in (2.5). As an example, given a hypothesis class $\mathcal{H}$, a distribution $\mathcal{D}$ over $\mathcal{Z}$, which is the space of data points, and a loss function $\ell : \mathcal{H} \times \mathcal{Z} \mapsto \mathbb{R}$, a supervised learning subtask aims at minimizing the risk $\mathbb{E}_{z\sim\mathcal{D}}[\ell(h,z)]$ over $h \in \mathcal{H}$. We denote the supervised learning subtask $\mathcal{T}$ by the tuple $(\mathcal{D}, \ell, \mathcal{H})$. Similarly, an algorithm that solves supervised learning subtasks maps from $\mathcal{T} = (\mathcal{D}, \ell, \mathcal{H})$ to a hypothesis $h \in \mathcal{H}$, aiming at minimizing the risk $R(h) = \mathbb{E}_{z\sim\mathcal{D}}[\ell(h,z)]$ over $h \in \mathcal{H}$. In what follows, we denote by $H_\mathcal{T}$ the objective of a learning subtask $\mathcal{T}$. If $\mathcal{T}$ is a reinforcement learning subtask, we have $H_\mathcal{T} = J$, and if $\mathcal{T}$ is a supervised learning subtask, we have $H_\mathcal{T} = R$.

The goal of the meta-learner is to find $\theta^* \in \Theta$ that optimizes the population version of the meta-objective $\overline{L}(\theta)$, which is defined as

$$\overline{L}(\theta) = \mathbb{E}_{\mathcal{T}\sim\iota}\Big[H_\mathcal{T}\big(\mathscr{A}_\theta(\mathcal{T})\big)\Big]. \tag{2.6}$$

To approximately optimize $\overline{L}$ defined in (2.6) based on the sample $\{\mathcal{T}_i\}_{i\in[n]}$ of subtasks, the meta-learner optimizes the following empirical version of the meta-objective,

$$L(\theta) = \frac{1}{n} \cdot \sum_{i=1}^{n} H_{\mathcal{T}_i}\big(\mathscr{A}_\theta(\mathcal{T}_i)\big). \tag{2.7}$$

The algorithm $\mathscr{A}_{\theta^*}$ corresponding to the global optimum θ^* of (2.7) incorporates the past experience through the observed learning subtasks $\{\mathcal{T}_i\}_{i\in[n]}$, and therefore, facilitates the learning of a new subtask (Pentina & Lampert,

2014; Finn et al., 2017a; Amit & Meir, 2017; Yoon et al., 2018). As an example, in model-agnostic meta-learning (MAML) (Finn et al., 2017a) for supervised learning, the hypothesis class $\mathcal{H}$ is parameterized by h_θ with $\theta \in \Theta$, and the algorithm $\mathscr{A}_\theta$ performs one step of gradient descent with $\theta \in \Theta$ as the starting point. In this setting, MAML aims to find the globally optimal starting point θ^* by minimizing the following meta-objective by gradient descent,

$$L(\theta) = \frac{1}{n} \cdot \sum_{i=1}^{n} R_i\big(h_{\theta - \eta \cdot \nabla_\theta R_i(h_\theta)}\big),$$

where η is the learning rate of $\mathscr{A}_\theta$ and $R_i(h) = \mathbb{E}_{z \sim \mathcal{D}_i}[\ell(h, z)]$ is the risk of the supervised learning subtask $\mathcal{T}_i = (\mathcal{D}_i, \ell, \mathcal{H})$. Similarly, in MAML for reinforcement learning, the algorithm $\mathscr{A}_\theta$ performs, e.g., one step of policy gradient with θ as the starting point. We call π_θ the main effect in the sequel. MAML aims to find the globally optimal main effect π_{θ^*} by maximizing the following meta-objective by gradient ascent,

$$L(\theta) = \frac{1}{n} \cdot \sum_{i=1}^{n} J_i\big(\pi_{\theta + \eta \cdot \nabla_\theta J_i(\pi_\theta)}\big),$$

where η is the learning rate of $\mathscr{A}_\theta$ and J_i is the expected total reward of the reinforcement learning subtask $\mathcal{T}_i = (\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i)$.

3. Meta-Reinforcement Learning

In this section, we present the analysis of meta-reinforcement learning (meta-RL). We first define the detailed problem setup of meta-RL and propose a meta-RL algorithm. We then characterize the global optimality of the stationary point attained by such an algorithm. We refer the analysis of meta-RL with neural network parameterization to §A.2.

3.1. Problem Setup and Algorithm

In meta-RL, the meta-learner observes a sample of MDPs $\{(\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i)\}_{i \in [n]}$ drawn independently from a task distribution ι. We set the algorithm $\mathscr{A}_\theta$ that optimizes the policy to be one step of (a variant of) proximal policy optimization (PPO) (Schulman et al., 2015; 2017) starting from the main effect π_θ. More specifically, $\mathscr{A}_\theta$ solves the following maximization problem,

$$\begin{aligned} &\mathscr{A}_\theta(\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i) \\ &\quad = \operatorname*{argmax}_{\pi} \mathbb{E}_{s \sim \nu_{i, \pi_\theta}} \Big[\langle Q_i^{\pi_\theta}(s, \cdot), \pi(\cdot \,|\, s) \rangle \\ &\qquad - 1/\eta \cdot D_{\mathrm{KL}}\big(\pi(\cdot \,|\, s) \big\| \pi_\theta(\cdot \,|\, s)\big) \Big]. \end{aligned} \tag{3.1}$$

Here $\langle \cdot, \cdot \rangle$ is the inner product over $\mathbb{R}^{|\mathcal{A}|}$, η is the tuning parameter of $\mathscr{A}_\theta$, and $Q_i^{\pi_\theta}$, ν_{i, π_θ} are the action-value function and the state visitation measure, respectively, corresponding to the MDP $(\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i)$ and the policy π_θ. Note that the objective in (3.1) has $D_{\mathrm{KL}}(\pi(\cdot \,|\, s) \| \pi_\theta(\cdot \,|\, s))$ in place of $D_{\mathrm{KL}}(\pi_\theta(\cdot \,|\, s) \| \pi(\cdot \,|\, s))$ compared with the original version of PPO (Schulman et al., 2015; 2017). As shown by (Liu et al., 2019), such a variant of PPO enjoys global optimality and convergence.

We parameterize the main effect π_θ as the following energy-based policy (Haarnoja et al., 2017) for all $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$\pi_\theta(a \,|\, s) = \frac{\exp\big(1/\tau \cdot \phi(s, a)^\top \theta\big)}{\sum_{a' \in \mathcal{A}} \exp\big(1/\tau \cdot \phi(s, a')^\top \theta\big)}, \tag{3.2}$$

where $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$ is the feature mapping, $\theta \in \mathbb{R}^d$ is the parameter, $\phi(\cdot, \cdot)^\top \theta$ is the energy function, and τ is the temperature parameter. The maximizer $\pi_{i, \theta} = \mathscr{A}_\theta(\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i)$ defined in (3.1) then takes the following form (Liu et al., 2019, Proposition 3.1) for all $s \in \mathcal{S}$,

$$\pi_{i, \theta}(\cdot \,|\, s) \propto \exp\big(1/\tau \cdot \phi(s, \cdot)^\top \theta + \eta \cdot Q_i^{\pi_\theta}(s, \cdot)\big). \tag{3.3}$$

The goal of meta-RL is to find the globally optimal main effect π_θ by maximizing the following meta-objective,

$$\begin{aligned} L(\theta) &= \frac{1}{n} \cdot \sum_{i=1}^{n} J_i(\pi_{i, \theta}), \\ &\text{where } \pi_{i, \theta} = \mathscr{A}_\theta(\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i). \end{aligned} \tag{3.4}$$

Here J_i is the expected total reward defined in (2.5) corresponding to the MDP $(\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i)$ for all $i \in [n]$. To maximize $L(\theta)$, we use gradient ascent, which iteratively updates θ as follows,

$$\theta_{\ell+1} \leftarrow \theta_\ell + \alpha_\ell \cdot \nabla_\theta L(\theta_\ell), \quad \text{for } \ell = 0, 1, \ldots, T-1, \tag{3.5}$$

where $\nabla_\theta L(\theta_\ell)$ is the gradient of the meta-objective at θ_ℓ, α_ℓ is the learning rate at the ℓ-th iteration, and T is the number of iterations. It remains to calculate the gradient $\nabla_\theta L(\theta)$. To this end, we first define the state-action vis-

itation measures induced by the main effect π_θ, and then calculate $\nabla_\theta L(\theta)$ in closed form based on such state-action visitation measures.

Definition 3.1 (Visitation Measures of Main Effect)**.** For all $i \in [n]$, given the MDP $(\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i)$ and the main effect π_θ, we denote by σ_{i,π_θ} the state-action visitation measure induced by the main effect π_θ. We further define the state-action visitation measure $\sigma_{i,\pi_\theta}^{(s,a)}$ initialized at $(s,a) \in \mathcal{S} \times \mathcal{A}$ as follows,

$$\sigma_{i,\pi_\theta}^{(s,a)}(s',a') = (1-\gamma_i) \cdot \sum_{t=0}^{\infty} \gamma_i^t \cdot \mathbb{P}(s_t = s', a_t = a'), \tag{3.6}$$

where $(s',a') \in \mathcal{S} \times \mathcal{A}$, $s_0 \sim P_i(\cdot \,|\, s,a)$, $s_{t+1} \sim P_i(\cdot \,|\, s_t, a_t)$, and $a_t \sim \pi_\theta(\cdot \,|\, s_t)$ for all $t \geq 0$.

In other words, given the transition kernel P_i and the discount factor γ_i, $\sigma_{i,\pi_\theta}^{(s,a)}$ is the state-action visitation measure induced by the main effect π_θ where the initial state distribution is given by $s_0 \sim P_i(\cdot \,|\, s,a)$. Based on the policy gradient theorem (Sutton & Barto, 2018), the following proposition calculates the gradient of the meta-objective L defined in (3.4) with respect to the parameter θ of the main effect π_θ.

Proposition 3.2 (Gradient of Meta-Objective)**.** It holds for all $\theta \in \mathbb{R}^d$ that

$$\nabla_\theta L(\theta) = \frac{1}{n} \cdot \sum_{i=1}^{n} \mathbb{E}_{(s,a)\sim\sigma_{\pi_{i,\theta}}} \big[h_{i,\theta}(s,a) \cdot A_i^{\pi_{i,\theta}}(s,a) \big], \tag{3.7}$$

where the auxiliary function $h_{i,\theta}$ takes the form of

$$\begin{aligned} h_{i,\theta}(s,a) &= 1/\tau \cdot \phi(s,a) \\ &\quad + \eta \cdot \gamma_i/\tau \cdot \mathbb{E}_{(s',a')\sim\sigma_{i,\pi_\theta}^{(s,a)}} \big[\phi(s',a') \cdot A_i^{\pi_\theta}(s',a') \big]. \end{aligned} \tag{3.8}$$

Here $A_i^{\pi_{i,\theta}}$ and $A_i^{\pi_\theta}$ are the advantage functions of the policy $\pi_{i,\theta}$ and the main effect π_θ, respectively, both corresponding to the MDP $(\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i)$. Also, $\sigma_{i,\pi_\theta}^{(s,a)}$ is the state-action visitation measure induced by the main effect π_θ defined in Definition 3.1, and $\sigma_{\pi_{i,\theta}}$ is the state-action visitation measure induced by the policy $\pi_{i,\theta}$, both corresponding to the MDP $(\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i)$.

Proof. See §D.1 for a detailed proof. □

In the sequel, we assume without loss of generality that the action-value function Q^π is available once we obtain the policy π, and the expectations over state-action visitation measures in (3.7) and (3.8) of Theorem 3.2 are available once we obtain the policies $\{\pi_{i,\theta}\}_{i\in[n]}$ and the main effect π_θ. We summarize meta-RL in Algorithm 1. In practice, we can estimate the action-value functions by temporal difference learning (Sutton, 1988) and the expectations over the visitation measures by Monte Carlo sampling (Konda, 2002).

Algorithm 1 Meta-RL

Require: Sampled MDPs $\{(\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i)\}_{i\in[n]}$ from the task distribution τ, feature mapping ϕ, number of iterations T, learning rate $\{\alpha_\ell\}_{\ell\in[T]}$, temperature parameter $1/\tau$, tuning parameter τ, initial parameter θ_0.

Initialization:
for $\ell = 0, \ldots, T-1$ **do**
  for $i \in [n]$ **do**
    Update the policy: $\pi_{i,\theta_\ell}(\cdot \,|\, s) \propto \exp\big(1/\tau \cdot \phi(s,\cdot)^\top \theta_\ell + \eta \cdot Q_i^{\pi_{\theta_\ell}}(s,\cdot)\big)$.
    Compute the auxiliary function $h_{i,\theta_\ell}(s,a)$ via (3.8)
  end for
  Compute the gradient of meta-objective $\nabla_\theta L(\theta_\ell)$ based on the policies $\{\pi_{i,\theta_\ell}\}_{i\in[n]}$ and auxiliary functions $\{h_{i,\theta_\ell}\}_{i\in[n]}$ via (3.7).
  Update the parameter of the main effect: $\theta_{\ell+1} \leftarrow \theta_\ell + \alpha_\ell \cdot \nabla_\theta L(\theta_\ell)$.
  Update the main effect: $\pi_{\theta_{\ell+1}}(\cdot \,|\, s) \propto \exp\big(1/\tau \cdot \phi(s,\cdot)^\top \theta_{\ell+1}\big)$.
end for
Output: θ_T and π_{θ_T}.

3.2. Theoretical Results

In this section, we analyze the global optimality of the ϵ-stationary point attained by meta-RL (Algorithm 1). In the sequel, we assume that the reward functions $\{r_i\}_{i\in[n]}$ are upper bounded by an absolute constant $Q_{\max} > 0$ in absolute value. It then follows from (2.1) and (2.2) that $|V_i^\pi(s,a)|$ and $|Q_i^\pi(s,a)|$ are upper bounded by $Q_{\max}$ for all $i \in [n]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$. Here we define Q_i^π and V_i^π as the state- and action-value functions of the policy π, respectively, corresponding to the MDP $(\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i)$.

To analyze the global optimality of meta-RL, we define the following meta-visitation measures induced by the main effect π_θ.

Definition 3.3 (Meta-Visitation Measures)**.** We define the joint meta-visitation measure ρ_{i,π_θ} over $(s',a',s,a) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathcal{A}$ induced by the main effect π_θ and the policy

$\pi_{i,\theta}$ as follows,

$$\rho_{i,\pi_\theta}(s', a', s, a) = \sigma_{i,\pi_\theta}^{(s,a)}(s', a') \cdot \sigma_{\pi_{i,\theta}}(s, a). \tag{3.9}$$

We further define the meta-visitation measure ς_{i,π_θ} as the marginal distribution of the joint meta-visitation measure ρ_{i,π_θ} of $(s', a') \in \mathcal{S} \times \mathcal{A}$, that is,

$$\varsigma_{i,\pi_\theta}(s', a') = \mathbb{E}_{(s,a)\sim\sigma_{\pi_{i,\theta}}}\big[\sigma_{i,\pi_\theta}^{(s,a)}(s', a')\big]. \tag{3.10}$$

In addition, for $(s', a') \in \mathcal{S} \times \mathcal{A}$ we define the mixed meta-visitation measure ϱ_{π_θ} over all the subtasks as follows,

$$\varrho_{\pi_\theta}(s', a') = \frac{1}{n} \cdot \sum_{i=1}^{n} \varsigma_{i,\pi_\theta}(s', a'). \tag{3.11}$$

In other words, the meta-visitation measure ς_{i,π_θ} is the state-action visitation measure induced by π_θ given the transition kernel P_i, the discount factor γ_i, and the initial state distribution $s_0 \sim \mathbb{E}_{(s,a)\sim\sigma_{\pi_{i,\theta}}}[P_i(\cdot \,|\, s, a)]$.

In what follows, we impose an assumption on the meta-visitation measures defined in Definition 3.3.

Assumption 3.4 (Regularity Condition on Meta-Visitation Measures)**.** We assume for all $\theta \in \mathbb{R}^d$ and $i \in [n]$ that

$$\mathbb{E}_{(s',a')\sim\varrho_{\pi_\theta}}\Big[\big(\mathrm{d}\sigma_{\pi_{i,\theta}}/\,\mathrm{d}\varrho_{\pi_\theta}(s', a')\big)^2\Big] \le C_0^2, \tag{3.12}$$

$$\mathbb{E}_{(s',a')\sim\varrho_{\pi_\theta}}\Big[\big(\mathrm{d}\varsigma_{i,\pi_\theta}/\,\mathrm{d}\varrho_{\pi_\theta}(s', a')\big)^2\Big] \le C_0^2, \tag{3.13}$$

where $C_0 > 0$ is an absolute constant . Here ς_{i,π_θ} and ϱ_{π_θ} are the meta-visitation measure and the mixed meta-visitation measure induced by the main effect π_θ, which are defined in (3.10) and (3.11) of Definition 3.3, respectively. Meanwhile, $\sigma_{\pi_{i,\theta}}$ is the state-action visitation measure induced by the policy $\pi_{i,\theta}$, which is defined in (2.4). Here $\mathrm{d}\sigma_{\pi_{i,\theta}}/\,\mathrm{d}\varrho_{\pi_\theta}$ and $\mathrm{d}\varsigma_{i,\pi_\theta}/\,\mathrm{d}\varrho_{\pi_\theta}$ are the Radon-Nikodym derivatives.

According to (3.11) of Definition 3.3, the upper bound in (3.12) of Assumption 3.4 holds if the $L_2(\varrho_{\pi_\theta})$-norms of $\mathrm{d}\sigma_{\pi_{i,\theta}}/\,\mathrm{d}\varsigma_{j,\pi_\theta}$ is upper bounded by C_0 for all $i, j \in [n]$. For $i = j$, note that $\pi_{i,\theta}$ is obtained by one step of PPO with π_θ as the starting point. Thus, for a sufficiently small tuning parameter η in (3.3), $\pi_{i,\theta}$ is close to π_θ. Hence, the assumption that $\mathrm{d}\sigma_{\pi_{i,\theta}}/\,\mathrm{d}\varsigma_{j,\pi_\theta}$ has an upper bounded $L_2(\varrho_{\pi_\theta})$-norm for all $i = j$ is a mild regularity condition. For $i \neq j$, to ensure the upper bound of the $L_2(\varrho_{\pi_\theta})$-norms of $\mathrm{d}\sigma_{\pi_{i,\theta}}/\,\mathrm{d}\varsigma_{j,\pi_\theta}$ in (3.12), Assumption 3.4 requires the task distribution ι to generate similar MDPs so that the meta-visitation measures $\{\varsigma_{i,\pi_\theta}\}_{i\in[n]}$ are similar across all the subtasks indexed by $i \in [n]$. Similarly, to ensure the upper bound in (3.13), Assumption 3.4 also requires that the meta-visitation measures $\{\varsigma_{i,\pi_\theta}\}_{i\in[n]}$ are similar across all the subtasks indexed by $i \in [n]$.

The following theorem characterizes the optimality gap of the ϵ-stationary point attained by meta-RL (Algorithm 1). Let θ^* be a global maximizer of the meta-objective $L(\theta)$ defined in (3.4). For all $(s', a') \in \mathcal{S} \times \mathcal{A}$ and $\omega \in \mathbb{R}^d$, we define

$$f_\omega(s', a') = \bigg(\sum_{i=1}^{n} \frac{A_i^{\pi_{i,\omega}}(s', a')}{1 - \gamma_i} \cdot \frac{\mathrm{d}\sigma_{\pi_{i,\theta^*}}}{\mathrm{d}\varrho_{\pi_\omega}}(s', a')\bigg) \Big/ \bigg(\sum_{i=1}^{n} g_{i,\omega}(s', a') \cdot \frac{\mathrm{d}\varsigma_{i,\pi_\omega}}{\mathrm{d}\varrho_{\pi_\omega}}(s', a')\bigg), \tag{3.14}$$

where we defined $g_{i,\omega}$ as follows,

$$g_{i,\omega}(s', a') = 1/\tau \cdot A_i^{\pi_{i,\omega}}(s', a') \cdot (\,\mathrm{d}\sigma_{\pi_{i,\omega}}/\,\mathrm{d}\varsigma_{i,\pi_\omega})(s', a') + \gamma_i \cdot \eta/\tau \cdot G_{i,\pi_\omega}(s', a') \cdot A_i^{\pi_\omega}(s', a').$$

Here τ is the temperature parameter in (3.2), η is the tuning parameter defined in (3.1), $A_i^{\pi_{i,\omega}}$ and $A_i^{\pi_\omega}$ are the advantage functions of the policy $\pi_{i,\omega}$ and the main effect π_ω, respectively, corresponding to the MDP $(\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i)$, and G_{i,π_ω} is defined as follows,

$$G_{i,\pi_\omega}(s', a') = \mathbb{E}_{(s',a',s,a)\sim\rho_{i,\pi_\omega}}\big[A_i^{\pi_{i,\omega}}(s, a) \,\big|\, s', a'\big], \tag{3.15}$$

where ρ_{i,π_ω} is the joint meta-visitation measure defined in (3.9) of Definition 3.3.

Theorem 3.5 (Optimality Gap of ϵ-Stationary Point)**.** Under Assumption 3.4, for all $R > 0$, $\omega \in \mathbb{R}^d$, and $\epsilon > 0$ such that

$$\nabla_\omega L(\omega)^\top v \le \epsilon, \quad \forall v \in \mathcal{B} = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le 1\}, \tag{3.16}$$

we have

$$L(\theta^*) - L(\omega) \le R \cdot \epsilon + C \cdot \inf_{v\in\mathcal{B}_R} \|f_\omega(\cdot, \cdot) - \phi(\cdot, \cdot)^\top v\|_{\varrho_{\pi_\omega}}, \tag{3.17}$$

where $\mathcal{B}_R = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le R\}$, $\overline{\gamma} = (\sum_{i=1}^n \gamma_i)/n$, and $C = 2C_0 \cdot Q_{\max}/\tau \cdot (1 + 2Q_{\max} \cdot \overline{\gamma} \cdot \eta)$. Here C_0 is defined in Assumption 3.4, τ is the temperature parameter

in (3.2), η is the tuning parameter defined in (3.1), and $Q_{\max}$ is the upper bound of the reward functions $\{r_i\}_{i\in[n]}$ in absolute value.

Proof. See §C.1 for a detailed proof. □

By Theorem 3.5, the global optimality of the ϵ-stationary point ω hinges on the representation power of the linear class $\{\phi(\cdot)^\top\theta : \theta \in \mathcal{B}_R\}$. More specifically, if the function f_ω defined in (3.14) is well approximated by $\phi(\cdot)^\top\theta$ for a parameter $\theta \in \mathcal{B}_R$, then ω is approximately globally optimal.

4. Meta-Supervised Learning

In this section, we present the analysis of meta-supervised learning (meta-SL). We first define the detailed problem setup of meta-SL and present a meta-SL algorithm. We then characterize the global optimality of the stationary point attained by such an algorithm. We refer the analysis of meta-SL with neural network parameterization to §A.3.

4.1. Problem Setup and Algorithm

In meta-SL, the meta-learner observes a sample of supervised learning subtasks $\{(\mathcal{D}_i, \ell, \mathcal{H})\}_{i\in[n]}$ drawn independently from a task distribution ι. Specifically, each subtask $(\mathcal{D}_i, \ell, \mathcal{H})$ consists of a distribution $\mathcal{D}_i$ over $\mathcal{X}\times\mathcal{Y}$, where $\mathcal{Y}\subseteq\mathbb{R}$, a loss function $\ell : \mathcal{H}\times\mathcal{X}\times\mathcal{Y}\mapsto\mathbb{R}$, and a hypothesis class $\mathcal{H}$. Each hypothesis $h\in\mathcal{H}$ is a mapping from $\mathcal{X}$ to $\mathcal{Y}$. The goal of the supervised learning subtask $(\mathcal{D}_i, \ell, \mathcal{H})$ is to obtain the following hypothesis,

$$h_i^* = \operatorname*{argmin}_{h\in\mathcal{H}} R_i(h) = \operatorname*{argmin}_{h\in\mathcal{H}} \mathbb{E}_{z\sim\mathcal{D}_i}\big[\ell(h, z)\big], \quad (4.1)$$

where $R_i(h) = \mathbb{E}_{z\sim\mathcal{D}_i}[\ell(h, z)]$ is the risk of $h\in\mathcal{H}$. To approximately attain the minimizer defined in (4.1), we parameterize the hypothesis class $\mathcal{H}$ by $\mathcal{H}_\theta$ with a feature mapping $\phi : \mathcal{X}\mapsto\mathbb{R}^d$ as follows,

$$\mathcal{H}_\theta = \big\{h_\theta(\cdot) = \phi(\cdot)^\top\theta : \theta\in\mathbb{R}^d\big\}, \quad (4.2)$$

and minimize $R_i(h_\theta)$ over $\theta\in\mathbb{R}^d$. We set the algorithm $\mathscr{A}_\theta$ in (2.7), which solves $(\mathcal{D}_i, \ell, \mathcal{H})$, to be one step of gradient descent with the starting point θ, that is,

$$\mathscr{A}_\theta(\mathcal{D}_i, \ell, \mathcal{H}) = h_{\theta-\eta\cdot\nabla_\theta R_i(h_\theta)}. \quad (4.3)$$

Here η is the learning rate of $\mathscr{A}_\theta$. The goal of meta-SL is to minimize the following meta-objective,

$$L(\theta) = \frac{1}{n}\cdot\sum_{i=1}^n R_i(h_{\theta_i}), \quad \text{where } h_{\theta_i} = \mathscr{A}_\theta(\mathcal{D}_i, \mathcal{R}, \mathcal{H}). \quad (4.4)$$

To minimize $L(\theta)$ defined in (4.4), we adopt gradient descent, which iteratively updates θ_ℓ as follows,

$$\theta_{\ell+1} \leftarrow \theta_\ell - \alpha_\ell\cdot\nabla_\theta L(\theta_\ell), \quad \text{for } \ell = 0, 1, \ldots, T-1. \quad (4.5)$$

Here $\nabla_\theta L(\theta_\ell)$ is the gradient of the meta-objective at θ_ℓ, α_ℓ is the learning rate at the ℓ-th iteration, and T is the number of iterations. (Fallah et al., 2019) show that the update defined in (4.5) converges to an ϵ-stationary point of the meta-objective L under a smoothness assumption on L. In what follows, we characterize the optimality gap of such an ϵ-stationary point.

We first introduce the Fréchet differentiability of the risk R_i in (4.1).

Definition 4.1 (Fréchet Differentiability)**.** Let $\mathcal{H}$ be a Banach space with the norm $\|\cdot\|_{\mathcal{H}}$. A functional $R : \mathcal{H}\mapsto\mathbb{R}$ is Fréchet differentiable at $h\in\mathcal{H}$ if it holds for a bounded linear operator $A : \mathcal{H}\mapsto\mathbb{R}$ that

$$\lim_{h_1\in\mathcal{H},\ \|h_1\|_{\mathcal{H}}\to 0} \frac{|R(h+h_1) - R(h) - A(h_1)|}{\|h_1\|_{\mathcal{H}}} \to 0. \quad (4.6)$$

We define A as the Fréchet derivative of R at $h\in\mathcal{H}$, and write

$$D_h R(\cdot) = A(\cdot). \quad (4.7)$$

In what follows, we assume that the hypothesis class $\mathcal{H}$ with the $L_2(\rho)$-inner product is a Hilbert space, where ρ is a distribution over $\mathcal{X}$. Thus, following from the definition of the Fréchet derivative in Definition 4.1 and the Rieze representation theorem (Rudin, 2006), it holds for an $a_h\in\mathcal{H}$ that

$$D_h R(\cdot) = A(\cdot) = \langle\cdot, a_h\rangle_{\mathcal{H}}, \quad (4.8)$$

Here we denote by $\langle f, g\rangle_{\mathcal{H}} = \int_{\mathcal{X}} f(x)\cdot g(x)\,\mathrm{d}\rho$ the $L_2(\rho)$-inner product. In what follows, we write

$$(\delta R/\delta h)(x) = a_h(x), \quad \forall x\in\mathcal{X}, h\in\mathcal{H}. \quad (4.9)$$

We refer to §B for an example of the Fréchet derivative defined in (4.9). We assume that $\mathcal{H}$ contains the parameterized hypothesis class $\mathcal{H}_\theta$ defined in (4.2), and impose the following assumption on the convexity and the Fréchet differentiability of the risk R_i in (4.1).

Assumption 4.2 (Convex and Differentiable Risk)**.** We assume for all $i \in [n]$ that the risk R_i defined in (4.1) is convex and Fréchet differentiable on $\mathcal{H}$.

Assumption 4.2 is a mild regularity condition on the risk R_i, which holds for the risks induced by commonly used loss functions, such as the squared loss and the cross entropy loss. Specifically, the convexity of R_i holds if the loss function $\ell(h, z)$ is convex in $h \in \mathcal{H}$ for all $z \in \mathcal{Z}$ (Rockafellar, 1968).

The following proposition holds under Assumption 4.2.

Proposition 4.3 (Convex and Differentiable Risk (Ekeland & Temam, 1999))**.** Under Assumption 4.2, it holds for all $i \in [n]$ that

$$R_i(h_1) \geq R_i(h_2) + \langle \delta R_i / \delta h_2, h_1 - h_2 \rangle_{\mathcal{H}}, \quad \forall h_1, h_2 \in \mathcal{H}.$$

Proof. See (Ekeland & Temam, 1999) for a detailed proof. □

We highlight that the convexity of the risks over $h \in \mathcal{H}$ does not imply the convexity of the meta-objective defined in (4.4). In contrast, Proposition 4.3 characterizes the functional geometry of the risk R_i in the Hilbert space $\mathcal{H}$ for all $i \in [n]$, which allows us to analyze the global optimality of meta-SL in the sequel.

4.2. Theoretical Results

In this section, we characterize the global optimality of the ϵ-stationary point attained by meta-SL defined in (4.5). Let θ^* be a global minimizer of the meta-objective $L(\theta)$ defined in (4.4), and ω be the ϵ-stationary point attained by meta-SL such that

$$\nabla_\omega L(\omega)^\top v \leq \epsilon, \quad \forall v \in \mathcal{B} = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq 1\}. \tag{4.10}$$

Our goal is to upper bound the optimality gap $L(\omega) - L(\theta^*)$. To this end, we first define the mixed distribution $\mathcal{M}$ over all the distributions $\{\mathcal{D}_i\}_{i\in[n]}$ as follows,

$$\mathcal{M}(x, y) = \frac{1}{n} \cdot \sum_{i=1}^{n} \mathcal{D}_i(x, y), \quad \forall (x, y) \in \mathcal{X} \times \mathcal{Y}. \tag{4.11}$$

To simplify the notation, we write ω_i and θ_i^* as the parameters that correspond to the outputs of the algorithms $\mathscr{A}_\omega(\mathcal{D}_i, \ell, \mathcal{H})$ and $\mathscr{A}_{\theta^*}(\mathcal{D}_i, \ell, \mathcal{H})$, respectively. More specifically, according to (4.3), we have

$$\begin{aligned} \omega_i &= \omega - \eta \cdot \nabla_\omega R_i(h_\omega), \\ \theta_i^* &= \theta^* - \eta \cdot \nabla_{\theta^*} R_i(h_{\theta^*}), \quad \forall i \in [n], \end{aligned} \tag{4.12}$$

where η is the learning rate of the algorithms $\mathscr{A}_\omega(\mathcal{D}_i, \ell, \mathcal{H})$ and $\mathscr{A}_{\theta^*}(\mathcal{D}_i, \ell, \mathcal{H})$.

The following theorem characterizes the optimality gap of the ϵ-stationary point ω attained by meta-SL. We define for all $(x, y, x') \in \mathcal{X} \times \mathcal{Y} \times \mathcal{X}$ that

$$w(x, y, x') = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{\delta R_i}{\delta h_{\omega_i}}(x') \cdot \frac{\mathrm{d}\mathcal{D}_i}{\mathrm{d}\mathcal{M}}(x, y), \tag{4.13}$$

$$u(x, y, x') = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{\delta R_i}{\delta h_{\omega_i}}(x') \cdot \frac{h_{\omega_i}(x') - h_{\theta_i^*}(x')}{w(x, y, x')}, \tag{4.14}$$

$$\phi_{\ell,\omega}(x, y, x') = \Big(I_d - \eta \cdot \nabla_\omega^2 \ell\big(\phi(x)^\top \omega, (x, y)\big)\Big)\phi(x'), \tag{4.15}$$

where $\mathrm{d}\mathcal{D}_i / \mathrm{d}\mathcal{M}$ is the Radon-Nikodym derivative and $\delta R_i / \delta h_{\omega_i}$ is the Fréchet derivative defined in (4.9).

Theorem 4.4 (Optimality Gap of ϵ-Stationary Point)**.** Let θ^* be a global minimizer of $L(\theta)$. Also, let ω be the ϵ-stationary point defined in (4.10). Let $\ell(h_\theta(x), (x, y))$ be twice differentiable with respect to all $\theta \in \mathbb{R}^d$ and $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Under Assumption 4.2, it holds for all $R > 0$ that

$$\begin{aligned} &L(\omega) - L(\theta^*) \\ &\leq \underbrace{R \cdot \epsilon}_{\text{(i)}} + \underbrace{\|w\|_{\mathcal{M}\cdot\rho}}_{\text{(ii)}} \cdot \underbrace{\inf_{v \in \mathcal{B}_R} \|u(\cdot) - \phi_{\ell,\omega}(\cdot)^\top v\|_{\mathcal{M}\cdot\rho}}_{\text{(iii)}}, \end{aligned} \tag{4.16}$$

where we define $\mathcal{B}_R = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq R\}$ as the ball with radius R and

$$\|w\|_{\mathcal{M}\cdot\rho} = \bigg(\int w^2(x, y, x') \,\mathrm{d}\mathcal{M}(x, y) \,\mathrm{d}\rho(x')\bigg)^{1/2}$$

as the $L_2(\mathcal{M} \cdot \rho)$-norm of w.

Proof. See §C.2 for a detailed proof. □

By Theorem 4.4, the optimality gap of the ϵ-stationary point ω hinges on the three terms on the right-hand side of (4.16). Here term (i) characterizes the deviation of the ϵ-stationary

point ω from a stationary point. Term (ii) characterizes the difficulty of all the subtasks sampled from the task distribution ι. Specifically, given the ϵ-stationary point ω, if the output h_{ω_i} of $\mathscr{A}_\omega(\mathcal{D}_i, \ell, \mathcal{H})$ well approximates the minimizer of the risk R_i in (4.1), then the Fréchet derivative $\delta R_i / \delta h_{\omega_i}$ defined in (4.9) is close to zero. Meanwhile, the Radon-Nikodym derivative $\mathrm{d}\mathcal{D}_i / \mathrm{d}\mathcal{M}$ characterizes the deviation of the distribution $\mathcal{D}_i$ from the mixed distribution $\mathcal{M}$ defined in (4.11), which is upper bounded if $\mathcal{D}_i$ is close to $\mathcal{M}$. Thus, term (ii) is upper bounded if h_{ω_i} well approximates the minimizer of R_i and $\mathcal{D}_i$ is close to $\mathcal{M}$ for all $i \in [n]$. Term (iii) characterizes the representation power of the feature mapping $\phi_{\ell,\omega}$ defined in (4.15). Specifically, if the function u defined in (4.14) of Theorem 4.4 is well approximated by $\phi_{\ell,\omega}(\cdot)^\top v$ for some $v \in \mathcal{B}_R$, then term (iii) is small. In conclusion, if the subtasks generated by the task distribution ι are sufficiently regular so that term (ii) is upper bounded, and the linear class $\{\phi_{\ell,\omega}(\cdot)^\top v : v \in \mathcal{B}_R\}$ has sufficient representation power, then ω is approximately globally optimal. See §B for a corollary of Theorem 4.4 when it is adapted to the squared loss.

References

Allen-Zhu, Z., Li, Y., and Liang, Y. Learning and generalization in overparameterized neural networks, going beyond two layers. *arXiv preprint arXiv:1811.04918*, 2018a.

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*, 2018b.

Amit, R. and Meir, R. Meta-learning by adjusting priors based on extended PAC-Bayes theory. *arXiv preprint arXiv:1711.01244*, 2017.

Arora, S., Du, S. S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584*, 2019.

Bai, Y. and Lee, J. D. Beyond linearization: On quadratic and higher-order approximation of wide neural networks. *arXiv preprint arXiv:1910.01619*, 2019.

Cai, Q., Yang, Z., Lee, J. D., and Wang, Z. Neural temporal-difference learning converges to global optima. In *Advances in Neural Information Processing Systems*, 2019.

Cao, Y. and Gu, Q. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *arXiv preprint arXiv:1905.13210*, 2019a.

Cao, Y. and Gu, Q. A generalization theory of gradient descent for learning over-parameterized deep ReLU networks. *arXiv preprint arXiv:1902.01384*, 2019b.

Chizat, L. and Bach, F. A note on lazy training in supervised differentiable programming. *arXiv preprint arXiv:1812.07956*, 2018.

Daniely, A. SGD learns the conjugate kernel class of the network. In *Advances in Neural Information Processing Systems*, 2017.

Du, S. S., Lee, J. D., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*, 2018a.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018b.

Ekeland, I. and Temam, R. *Convex analysis and variational problems*, volume 28. SIAM, 1999.

Evgeniou, T. and Pontil, M. Regularized multi–task learning. In *International Conference on Knowledge Discovery and Data Mining*, pp. 109–117, 2004.

Facchinei, F. and Pang, J.-S. *Finite-dimensional variational inequalities and complementarity problems*. Springer Science & Business Media, 2007.

Fallah, A., Mokhtari, A., and Ozdaglar, A. On the convergence theory of gradient-based model-agnostic meta-learning algorithms. *arXiv preprint arXiv:1908.10400*, 2019.

Fan, J., Ma, C., and Zhong, Y. A selective overview of deep learning. *arXiv preprint arXiv:1904.05526*, 2019.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, 2017a.

Finn, C., Yu, T., Zhang, T., Abbeel, P., and Levine, S. One-shot visual imitation learning via meta-learning. *arXiv preprint arXiv:1709.04905*, 2017b.

Finn, C., Xu, K., and Levine, S. Probabilistic model-agnostic meta-learning. In *Advances in Neural Information Processing Systems*, 2018.

Finn, C., Rajeswaran, A., Kakade, S., and Levine, S. Online meta-learning. *arXiv preprint arXiv:1902.08438*, 2019.

Gupta, A., Mendonca, R., Liu, Y., Abbeel, P., and Levine, S. Meta-reinforcement learning of structured exploration strategies. In *Advances in Neural Information Processing Systems*, 2018.

Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, 2017.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, 2002.

Khodak, M., Balcan, M.-F., and Talwalkar, A. Provable guarantees for gradient-based meta-learning. *arXiv preprint arXiv:1902.10644*, 2019.

Konda, V. *Actor-Critic Algorithms*. PhD thesis, Massachusetts Institute of Technology, 2002.

Lee, J., Xiao, L., Schoenholz, S. S., Bahri, Y., Sohl-Dickstein, J., and Pennington, J. Wide neural networks of any depth evolve as linear models under gradient descent. *arXiv preprint arXiv:1902.06720*, 2019.

Li, Y. and Liang, Y. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, 2018.

Li, Z., Zhou, F., Chen, F., and Li, H. Meta-SGD: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*, 2017.

Liu, B., Cai, Q., Yang, Z., and Wang, Z. Neural proximal/trust region policy optimization attains globally optimal policy. *arXiv preprint arXiv:1906.10306*, 2019.

Mendonca, R., Gupta, A., Kralev, R., Abbeel, P., Levine, S., and Finn, C. Guided meta-policy search. In *Advances in Neural Information Processing Systems*, 2019.

Nagabandi, A., Finn, C., and Levine, S. Deep online learning via meta-learning: Continual adaptation for model-based RL. *arXiv preprint arXiv:1812.07671*, 2018.

Nichol, A. and Schulman, J. Reptile: A scalable meta-learning algorithm. *arXiv preprint arXiv:1803.02999*, 2: 2, 2018.

Nichol, A., Achiam, J., and Schulman, J. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

Pan, S. J. and Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10): 1345–1359, 2009.

Pentina, A. and Lampert, C. A PAC-Bayesian bound for lifelong learning. In *International Conference on Machine Learning*, 2014.

Rajeswaran, A., Finn, C., Kakade, S. M., and Levine, S. Meta-learning with implicit gradients. In *Advances in Neural Information Processing Systems*, 2019.

Rakelly, K., Shelhamer, E., Darrell, T., Efros, A. A., and Levine, S. Few-shot segmentation propagation with guided networks. *arXiv preprint arXiv:1806.07373*, 2018.

Rockafellar, R. Integrals which are convex functionals. *Pacific journal of mathematics*, 24(3):525–539, 1968.

Rudin, W. *Real and complex analysis*. McGraw-Hill, 2006.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International Conference on Machine Learning*, 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Sutton, R. S. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Taylor, M. E. and Stone, P. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(Jul):1633–1685, 2009.

Thrun, S. and Pratt, L. *Learning to learn*. Springer Science & Business Media, 2012.

Wang, H., Sun, R., and Li, B. Global convergence and induced kernels of gradient-based meta-learning with neural nets. *arXiv preprint arXiv: 2006.14606*, 2020.

Weiss, K., Khoshgoftaar, T. M., and Wang, D. A survey of transfer learning. *Journal of Big Data*, 3(1):9, 2016.

Wu, L., Ma, C., and Weinan, E. How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. In *Advances in Neural Information Processing Systems*, 2018.

Xu, K., Ratner, E., Dragan, A., Levine, S., and Finn, C. Learning a prior over intent via meta-inverse reinforcement learning. *arXiv preprint arXiv:1805.12573*, 2018.

Yoon, J., Kim, T., Dia, O., Kim, S., Bengio, Y., and Ahn, S. Bayesian model-agnostic meta-learning. In *Advances in Neural Information Processing Systems*, 2018.

Yu, T., Abbeel, P., Levine, S., and Finn, C. One-shot hierarchical imitation learning of compound visuomotor tasks. *arXiv preprint arXiv:1810.11043*, 2018.

Zou, D., Cao, Y., Zhou, D., and Gu, Q. Stochastic gradient descent optimizes over-parameterized deep ReLU networks. *arXiv preprint arXiv:1811.08888*, 2018.